A SWN model for Multilevel Fat Tree

G. Franceschinis M. Beccuti

12th September 2004

1 Introduction

The model presented in this section, represents an interconnection network similar to that used in the *Connection Machine CM-5*. This type of interconnection network, introduced by Leiserson, is called *fat-tree*.

The net structure is a *k-ary* tree, whose leaves are the processors while the internal nodes are switching elements. The tree structure is called *fat* because the branches of the tree closer to the root are *"thicker"* (i.e. they may support a large number of communications in parallel).

The CM-5 data network is implemented with a 4-ary fat tree, in which each internal node is made up of several router chips. The model represents a very simple specification of the message workload: each processor p_i runs one process that performs local computation, then it sends a message to randomly chosen destination processor.

When a process mapped on processor i wants to communicate with a process, mapped on processor j, the message climbs up towards the root of the tree until the closest common ancestor is found, then it descends towards the destination processor leaf.

The network structure suggests a hierarchical naming schema for the tree nodes: each node is uniquely identified by a tuple, representing its path from the root. In this way it is easy to find the closest common ancestor of two nodes, which corresponds to the longest common prefix in their pathnames. This encoding easily translates in a specification for the color structure of the SWN model of the interconnection network: k-ary trees of depth D require a single (non ordered) color class C of cardinality k-1, and a token representing a node at depth level[1] j ($1 \leq j \leq D$) is a tuple of arity j. A greater optimization is to consider the color used to identify objects at different levels of the tree as independent. Infact if the components are not considered as independent by the SRG construction algorithm, the degree of aggregation achieved would be substantially decreased.

The description of the behavior of this system requires the definition of several parameters and operating rules:

[1] The depth level of the root is 0, while for the leaves it is D.

- **local computation time** LC**;**
- **communication time** M**;**
- **maximum number of concurrent communications through a node at depth** i N_i**;**
- **depth of tree** D**;**
- **arity of tree** k**.**

Our example will have a 4-ary($k = 4$) fat tree of depth $D = 3$, and the maximum number of concurrent communications through the switches at each level will be:

- **root switch** $= 8$**;**
- **first switch** $= 2$**;**
- **second switch** $= 4$**;**

The *local computation time* will be modelled by an exponential distribution with mean $\frac{1}{\lambda}$ and *communication time* by exponential distribution with mean $\frac{1}{\mu}$. The *probability of chosing a given destination* is a uniformly distributed among all remaining processors in the system.

Class id	**Description**	**Type**
C	tree node id	unordered
Subclass id	**Description**	**Elements**
C_1	subclass of C for identity of node at level 3	a, b, c, d
C_2	subclass of C for identity of node at level 2	a, b, c, d
C_3	subclass of C for identity of node at level 1	a, b, c, d

Table 1: Characteristics of color class

2 Description of the model

The SWN in Fig.1 represents a 4-ary fat-tree of depth three, and includes a very simple specification of the messages workload: each processor P runs one process, which performs local computation, then it sends a message to a randomly chosen destination (uniform distribution).

The three color classes: C_1, C_2, C_3, all of cardinality four, are used to represent the labels of the node at level $D - i + 1$ with $1 \leq i \leq 3$. In fact the processor i will be identified by the tuple (s_{i3},s_{i2},s_{i1}) with $s_{ij} \in C_j$. It is important to stress the fact that in order to fully exploit the symmetries

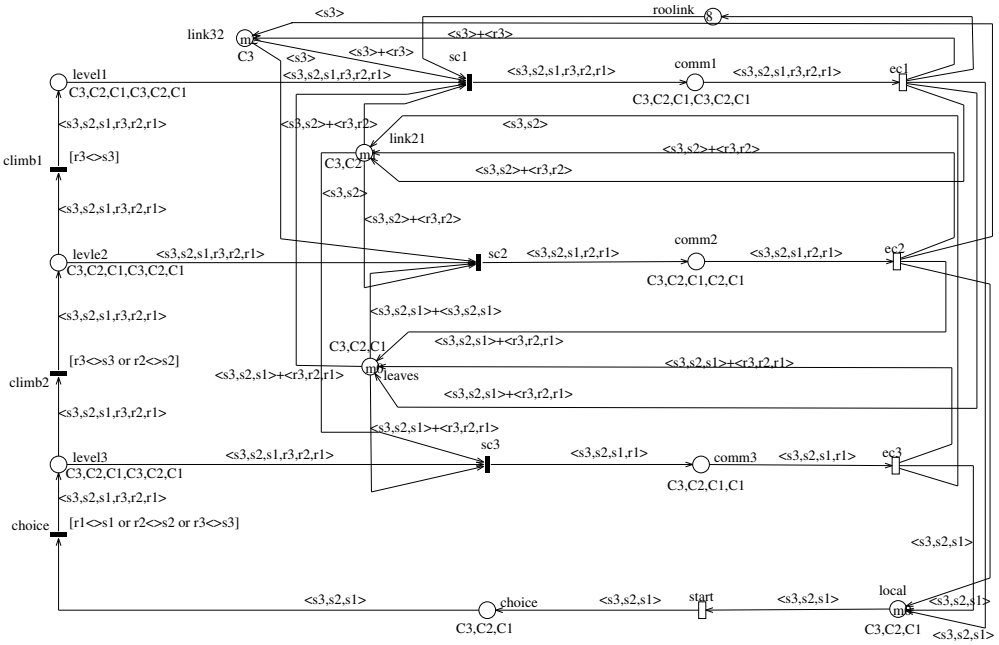

Figure 1: SWN model of the messages routing in a 3 level 4-ary fat-tree

through the SRG algorithm, the color used to identify nodes at different levels of the tree must be considered as independent.

The model can be divided in four sub-models: one representing the local computation, the others representing the subnet at level i, i=1,2,3.

Observe that any k-ary fat tree with three levels can be modeled by the same net structure, changing only the definition of the basic color class C. On the contrary, k-ary fat trees with $D \neq 3$ levels require to change the net structure.

2.1 Description of the sub-model of local computation

A token in place *local* represents the local computation, which ends with the firing of transition *start*. Immediate transition *choose* represents the choice of a destination for the message sent by the process.

2.2 Description of the sub-model of the generic i level

This sub-net contains two places $level_i$ and $comm_i$ and two transitions sc_i and ec_i,which simulate the communication between processes.

The transition $climb_J$[2] represents the search for the sender-receiver com-

[2]A predicate is associated with this transition that compares the prefixes of the two

mon ancestor.

The places *rootlink*(in the level 3 sub-net), *link21*(in the level 2 sub-net), *link32*(in the level 1-level), assure that the number of concurrent communications,which can pass through a switch at a given level i, can't exceed the maximum number of allowed concurrent communications N_i. In the example of Fig.1 the initial marking of place *rootlink* is 8, hence at most 8 communications, which need to be routed through the root, may be concurrently served. Place *link21* initially contains two copies of each identifier $<s_3, s_2> \in C_3 \times C_2$ meaning that at most two communications can be concurrently routed through the switches at level two. Finally, place *link32* initially contains four copies of each identifier $<s_3> \in C_3$ meaning that at most four communications can be concurrently routed through each switch at level one.

Place *leaves* assures that a process can send a message only to another process, which currently is not involved in any communicating. Summarizing, places *rootlink*, *link21*, *link32* and *leaves* simulate the resources needed to set up the communication between the sender and receiver processors.

Place	**Domain**
local	$C_3 \times C_2 \times C_1$
choice	$C_3 \times C_2 \times C_1$
$level_i$	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$
comm3	$C_3 \times C_2 \times C_1 \times C_1$
link21	$C_2 \times C_3$
comm2	$C_3 \times C_2 \times C_1 \times C_2 \times C_1$
link32	C_3
comm1	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$
rootlink	neutral

Table 2: Table of places

3 Size of SRG and RG

By analyzing the SRG and the RG sizes of this model we obtain an average reduction factor of about 3 by using the SRG instead of the RG technique; unfortunately the degree of reduction achieved is lower than the expected reduction. This is due to the fact that the specified color structure introduces an unnecessary dependence between nodes which have different father but the same associated color characterizing the postfix of the pathname.

This problem can be solved easily as follows. Define class C_i (representing the nodes at level d-i+1) as a color class of cardinality k^{d-i+1}, and

nodes up to the considered level

Transition	**Domain**	**Weight**
start	$C_3 \times C_2 \times C_1$	λ
ec3	$C_3 \times C_2 \times C_1 \times C_1$	μ
ec2	$C_3 \times C_2 \times C_1 \times C_2 \times C_1$	μ
ec1	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$	μ
choose	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$	1
sc3	$C_3 \times C_2 \times C_1 \times C_1$	1
climb2	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$	1
sc2	$C_3 \times C_2 \times C_1 \times C_2 \times C_1$	1
climb1	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$	1
sc1	$C_3 \times C_2 \times C_1 \times C_3 \times C_2 \times C_1$	1

Table 3: Table of transitions

then represent the father-children relation of the tree structure in the initial (symbolic) marking.

The Fig.2 depicts the new SWN model of a 3 level 4-ary tree, where the color classes have cardinality as follows: $|C1| = 64$, $|C2| = 16$ and $|C3| = 4$. Moreover the new places *conn21* and *conn32* represent respectively the father-children relation at the leaves level and at one level above.

This new representation allows us to achieve the goal of allowing independent permutations within the sets of the siblings associated with different fathers.This change in the color structure of the model allows us to achieve a reduction factor of about 4.5 by using the SRG instead of RG technique.

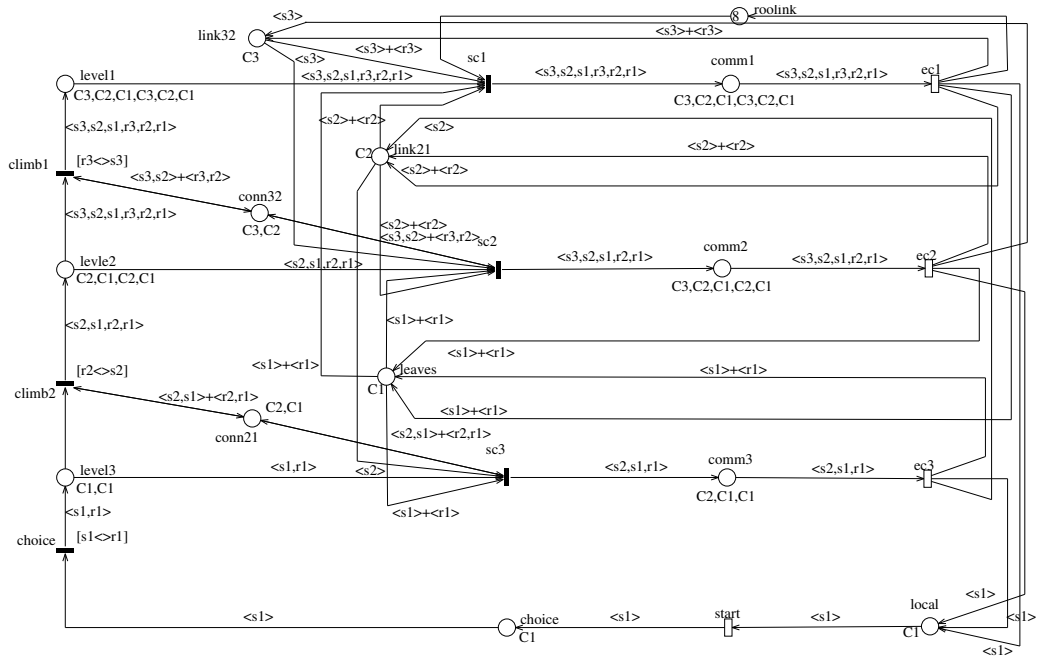

Figure 2: SWN model of 3 level 4-ary fat-tree interconnection capturing all permutation symmetries

References

[1] G. Chiola, G. Franceschinis and R. Gaeta. *Modeling Symmetric Computer Architectures by SWNs*. In PROC. 15th Intern. Conference on Application and Theory of Petri Nets, Zarogoza, Spain, June 1994. Lecture Notes in Computer Science, no. 815, Spinger-Verlag, 1994